데이터 마이닝과 퍼지인식도 기반의 인과관계 지식베이스 구축에 관한 연구

A Study on the Development of Causal Knowledge Base Based on Data Mining and Fuzzy Cognitive Map

김 진 성
전주대학교 경영학부 교수

Jin Sung Kim
School of Business Administration, Jeonju University
Hyoja-Dong 3-1200, Wansan-Ku, Jeonju, Jeonbuk, 560-759
kimjs@jeonju.ac.kr

Due to the increasing use of very large databases, mining useful information and implicit knowledge from databases is evolving. However, most conventional data mining algorithms identify the relationship among features using binary values (TRUE/FALSE or 0/1) and find simple IF-THEN rules at a single concept level. Therefore, implicit knowledge and causal relationships among features are commonly seen in real-world database and applications.

In this paper, we thus introduce the mechanism of mining fuzzy association rules and constructing causal knowledge base from database. A causal knowledge base construction algorithm based on Fuzzy Cognitive Map (FCM) and Srikant and Agrawal's association rule extraction method were proposed for extracting implicit causal knowledge from database. Fuzzy association rules are well suited for the thinking of human subjects and will help to increase the flexibility for supporting users in making decisions or designing the fuzzy systems. It integrates fuzzy set concept and causal knowledge-based data mining technologies to achieve this purpose.

The proposed mechanism consists of three phases: First, adaptation of the fuzzy membership function to the database. Second, extraction of the fuzzy association rules using fuzzy input values. Third, building the causal knowledge base. A credit example is presented to illustrate a detailed process for finding the fuzzy association rules from a specified database, demonstrating the effectiveness of the proposed algorithm.

Keywords: data mining, association rule, fuzzy membership functions, fuzzy cognitive map, causal knowledge base

1. Introduction

Data mining is one of hot topics in the field of knowledge discovery and management (Bonchi, et al., 2001; Chakrabarti et al., 1999; Changchien & Lu, 2001; Hui & Jha, 2000; Lee et al., 2002; Song et al., 2001). The association rule extraction mechanism, which was proposed by Agrawal et al.(1993), was a most popular tools to execute the data mining. Given a set of transactions, where each transaction is a set of item, an association rule is an expression of the form X → Y. X and Y means the sets of items. An example of an association rule is: "20% of transactions that contain beer also contain diapers; 10% of all transactions contain both these items." Here 20% is called the *confidence* of the rule, and 10% the *support* of the rule.

However, the most critical problem with data mining is the poor interpretability of mining results. In addition, interpreting the mining results is very difficult for general decision markers because they require high expertise in data mining or expert systems, etc (Lee et al., 2002). As a result, basic association rules couldn't represent the customer's implicit knowledge. In this sense, we propose a three-phased fuzzy and association rule-based causal knowledge based construction mechanism.

2. Methodology

Our proposed causal knowledge base construction mechanism was based on fuzzy membership function, association rule mining and fuzzy cognitive map (FCM). Which was aimed at enriching the reasoning ability and justification quality of knowledge based expert systems. The proposed mechanism consists of the three phases-fuzzy membership function, extraction of the fuzzy association rules, and development of the causal knowledge base. Figure 1 shows the research methodology.

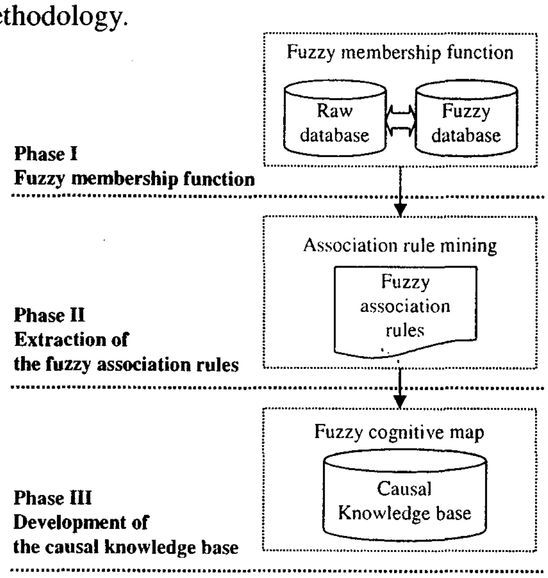

Figure 1. Research methodology

The first phase is to adapt the fuzzy membership function to traditional databases. In the second phase, we applied the association rules mining technique to extract the relationships among attributes. Final stage of the proposed hybrid causal knowledge base construction mechanism is to apply the FCM to the fuzzy association rules and construct the causal knowledge base.

3. Implementation

To prove the quality of hybrid causal knowledge base construction mechanism, we used credit data stored in U.C. Irvine's machine learning data repository. Totally 690 data was used for validation. Which was composed of 15 input variables and 1 output variable (attribute). The prototype system was implemented by using the Excel and VBA language in a Windows XP environment. In addition, SPSS and Clementine 6.0.1 was also used to preprocess the raw-data and extract the association rules. We call this prototype system as FAC (Fuzzy membership function and Association rule-based Causal knowledge base). Figure 2 shows the raw database for credit screening.

3.1 Phase 1: Fuzzy membership function

In the first phase, we adapted the fuzzy membership functions to transform the real data into fuzzy sets. Fuzzy membership functions used in this phase was as follows (Mitra & Pal, 1994):

$$\pi(F_j : c, \lambda) = \begin{Bmatrix} 2\left(1 - \dfrac{|F_j - c|}{\lambda}\right)^2, & \text{for } \dfrac{\lambda}{2} \le |F_j - c| \le \lambda \\ 1 - 2\left(\dfrac{|F_j - c|}{\lambda}\right)^2, & \text{for } 0 \le |F_j - c| \le \dfrac{\lambda}{2} \\ 0, & \text{otherwise} \end{Bmatrix}$$

Figure 3 shows the fuzzified datase transformed by fuzzy membership functions.

No	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16
1	b	30.83	0	u	g	w	v	1.25	t	t	1	f	g	202	0	+
2	a	58.67	4.46	u	g	q	h	3.04	t	t	6	f	g	43	560	+
3	a	24.5	0.5	u	g	q	h	1.5	t	f	0	f	g	280	824	+
4	b	27.83	1.54	u	g	w	v	3.75	t	t	5	t	g	100	3	+
5	b	20.17	5.625	u	g	w	v	1.71	t	f	0	f	s	120	0	+
6	b	32.08	4	u	g	m	v	2.5	t	f	0	t	g	360	0	+
7	b	33.17	1.04	u	g	r	h	6.5	t	f	0	t	g	164	31285	+
8	a	22.92	11.585	u	g	cc	v	0.04	t	f	0	f	g	80	1349	+
9	b	54.42	0.5	y	p	k	h	3.96	t	f	0	f	g	180	314	+
10	b	42.5	4.915	y	p	w	v	3.165	t	f	0	t	g	52	1442	+
11	b	22.08	0.83	u	g	c	h	2.165	f	f	0	t	g	128	0	+
12	b	29.92	1.835	u	g	c	h	4.335	t	f	0	f	g	260	200	+
13	a	38.25	6	u	g	k	v	1	t	f	0	t	g	0	0	+
14	b	48.08	6.04	u	g	k	v	0.04	f	f	0	f	g	0	2690	+
15	a	45.83	10.5	u	g	q	v	5	t	t	7	t	g	0	0	+
16	b	36.67	4.415	y	p	k	v	0.25	t	t	10	t	g	320	0	+
17	b	28.25	0.875	u	g	m	v	0.96	t	t	3	t	g	396	0	+
18	a	23.25	5.875	u	g	q	v	3.17	t	t	10	f	g	120	245	+
19	b	21.83	0.25	u	g	d	h	0.665	t	f	0	t	g	0	0	+
20	a	19.17	8.585	u	g	cc	h	0.75	t	t	7	f	g	96	0	+
21	b	25	11.25	u	g	c	v	2.5	t	t	17	f	g	200	1208	+
22	b	23.25	1	u	g	c	v	0.835	t	f	0	f	s	300	0	+
23	a	47.75	8	u	g	c	v	7.875	t	t	6	t	g	0	1260	+
24	a	27.42	14.5	u	g	x	h	3.085	t	t	1	f	g	120	11	+
25	a	41.17	6.5	u	g	q	v	0.5	t	t	3	t	g	145	0	+
26	a	15.83	0.585	u	g	c	h	1.5	t	t	2	f	g	100	0	+
27	a	47	13	u	g	i	bb	5.165	t	t	9	t	g	0	0	+
28	b	56.58	18.5	u	g	d	bb	15	t	t	17	t	g	0	0	+
29	b	57.42	8.5	u	g	e	h	7	t	t	3	f	g	0	0	+
30	b	42.08	1.04	u	g	w	v	5	t	t	6	t	g	500	10000	+

Figure 2. Raw database for credit screening

No	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16
1	b	low	low	u	g	w	v	low	t	t	low	f	g	low	low	up
2	a	high	low	u	g	q	h	low	t	t	low	f	g	low	low	up
3	a	low	low	u	g	q	h	low	t	f	low	f	g	low	low	up
4	b	low	low	u	g	w	v	low	t	t	low	t	g	low	low	up
5	b	low	low	u	g	w	v	low	t	f	low	f	s	low	low	up
6	b	low	low	u	g	m	v	low	t	f	low	t	g	low	low	up
7	b	low	low	u	g	r	h	low	t	f	low	t	g	low	low	up
8	a	low	medium	u	g	cc	v	low	t	f	low	f	g	low	low	up
9	b	high	low	y	p	k	h	low	t	f	low	f	g	low	low	up
10	b	medium	low	y	p	w	v	low	t	f	low	t	g	low	low	up
11	b	low	low	u	g	c	h	low	f	f	low	t	g	low	low	up
12	b	low	low	u	g	c	h	low	t	f	low	f	g	low	low	up
13	a	medium	low	u	g	k	v	low	t	f	low	t	g	low	low	up
14	b	medium	low	u	g	k	v	low	f	f	low	f	g	low	low	up
15	a	medium	medium	u	g	q	v	low	t	t	low	t	g	low	low	up
16	b	medium	low	y	p	k	v	low	t	t	low	t	g	low	low	up
17	b	low	low	u	g	m	v	low	t	t	low	t	g	low	low	up
18	a	low	low	u	g	q	v	low	t	t	low	f	g	low	low	up
19	b	low	low	u	g	d	h	low	t	f	low	t	g	low	low	up
20	a	low	low	u	g	cc	h	low	t	t	low	f	g	low	low	up
21	b	low	medium	u	g	c	v	low	t	t	low	f	g	low	low	up
22	b	low	low	u	g	c	v	low	t	f	low	f	s	low	low	up
23	a	medium	low	u	g	c	v	low	t	t	low	t	g	low	low	up
24	a	low	medium	u	g	x	h	low	t	t	low	f	g	low	low	up
25	a	medium	low	u	g	q	v	low	t	t	low	t	g	low	low	up
26	a	low	low	u	g	c	h	low	t	t	low	f	g	low	low	up
27	a	medium	medium	u	g	i	bb	low	t	t	low	t	g	low	low	up
28	b	high	high	u	g	d	bb	medium	t	t	low	t	g	low	low	up
29	b	high	low	u	g	e	h	low	t	t	low	f	g	low	low	up
30	b	medium	low	u	g	w	v	low	t	t	low	t	g	medium	low	up

Figure 3. Fuzzified database

3.2 Phase 2: Extraction of the fuzzy association rules

The association rule mining algorithm we adopted here is an APRIORI algorithm (Agrawal et al., 1993), which was known to yield a set of association rules. Based on the preprocessed credit database in Figure 3, the corresponding association rules were extracted with a threshold of 80% confidence. Table 1 shows an excerpt of the derived association rules. The association rules shown in Table 1 are straightforward and easy to understand and interpret. Figure 4 shows the association rule extraction process using Clementine.

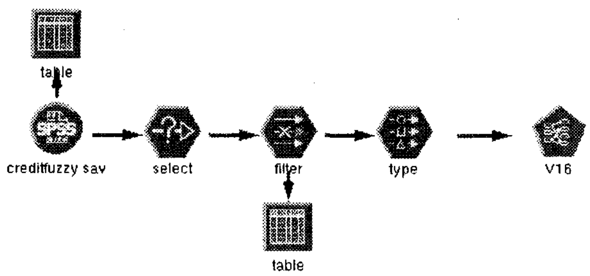

Figure 4. Association rule extraction process using Clementine

3.3 Phase 3: Development of the causal knowledge base

After the extraction of association rules, FCM-driven causal knowledge base was constructed. Figure 5(a,b) shows the FCM-driven causal knowledge base.

Table 2. Example of association rules from the fuzzified database

V15 == low <= V11 == low (334:99.405%, 0.991)
V11 == low <= V15 == low (333:99.107%, 0.994)
V8 == low <= V11 == low (334:99.405%, 0.943)
V11 == low <= V8 == low (316:94.048%, 0.997)
V14 == low <= V11 == low (334:99.405%, 0.922)
V11 == low <= V14 == low (310:92.262%, 0.994)
V13 == g <= V11 == low (334:99.405%, 0.913)
V11 == low <= V13 == g (307:91.369%, 0.993)
....
....
V2 == low <= V6 == w (35:10.417%, 0.8)
V16 == up <= V10 == t (177:52.679%, 0.819)
V7 == v <= V6 == w (35:10.417%, 0.886)
V10 == f <= V9 == f (92:27.381%, 0.826)
V16 == down <= V9 == f (92:27.381%, 0.848)

4. Conclusions

The result of experiment with an illustrative database proved to b valid and robust. In conclusion, this study has shown how the fuzzy association rules and FCM can be brought together to create causal knowledge base. It is expected that the proposed hybrid causal knowledge base construction mechanism will have a significant impact on the research domain related to the

	V1_b	V10_f	V11_low	V13_g	V14_low	V15_low	V16_down	V16_up	V2_low	V3_low	V4_u	V4_y	V5_g	V5_p	V7_v	V8_low	V9_t
V1_a			1.00	0.96	0.93	0.99										0.96	
V1_b			0.99	0.89	0.92	0.99				0.81						0.93	
V10_f			1.00	0.82	0.91	0.99				0.84						0.97	
V10_t			0.99	0.99	0.94	0.99		0.82			0.85		0.85			0.92	0.91
V11_low				0.91	0.92	0.99										0.94	
V12_f			0.99	0.92	0.97	0.99				0.80						0.96	
V12_t			0.99	0.91	0.87	0.99										0.92	0.80
V13_g			0.99		0.93	0.99										0.94	
V14_low			0.99	0.92		1.00										0.94	
V15_low			0.99	0.92	0.93											0.94	
V2_low			1.00	0.90	0.93	0.99				0.80						0.99	
V2_medium			1.00	0.92	0.86	0.99				0.86	0.81		0.81			0.87	0.83
V3_low			1.00	0.90	0.92	0.99										0.95	
V3_medium			0.98	0.98	0.97	1.00					0.87		0.87			0.93	0.85
V4_u			1.00	0.91	0.92	0.99							1.00			0.93	
V4_y			0.99	0.94	0.96	1.00				0.86				1.00		0.99	
V5_g			1.00	0.91	0.92	0.99					1.00					0.93	
V5_p			0.99	0.94	0.96	1.00				0.86		1.00				0.99	
V6_c			0.99	0.92	0.95	1.00										0.94	
V6_q			1.00	0.98	0.96	1.00				0.81	0.91		0.91			0.94	0.89
V6_w	0.83		1.00	0.91	0.91	1.00			0.80	0.83	0.83		0.83		0.89	0.94	
V7_h			1.00	0.96	0.92	1.00				0.81						0.94	0.81
V7_v			0.99	0.90	0.93	1.00										0.96	
V8_low			1.00	0.91	0.92	0.99				0.80							
V9_f		0.83	1.00	0.87	0.90	0.93	0.85		0.81	0.83						1.00	
V9_t			0.99	0.93	0.93	0.99		0.82			0.82		0.82			0.92	

Figure 5(a). FCM-driven causal knowledge base: matrix

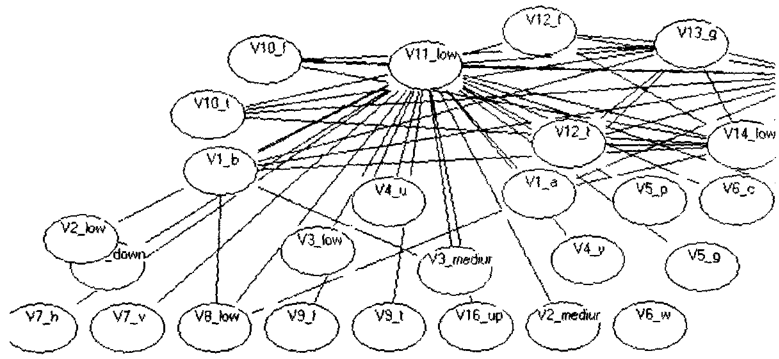

Figure 5(b). Sub example of FCM-driven causal knowledge base: graph

perception and knowledge management. Further research topics still remaining are as follows:

(1) The basic technology of association rule mining used for this study needs to be improved so that more fuzzy knowledge can be analyzed.

(2) Fuzzy membership functions need to be integrated with other rule refining and reasoning mechanism.

(3) FCM construction processes need to be improved with other useful knowledge management mechanisms.

References

[1] Agrawal, R., Imielinski T., and Swami, A., "Mining Association Rules between Sets of Items in large Databases," *In Proc. of the ACM SIGMOD Conference on Management of Data*, Washington, D.C., pp.207-216, 1993.

{2} Bonchi, F., Giannotti, F., Gozzi, C., Manco, G., Nanni, M., Pedreschi, D., Renso, C., and Ruggieri, S., "Web Log Data Warehousing and Mining for Intelligent Web Caching," *Data & Knowledge Engineering*, 39, pp.165-189, 2001.

[3] Chakrabarti, S., Dom, B.E., Kumar, S. R., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D., and Kleinberg, J.M., "Mining the Web's Link Structure," *Computer*, 32, pp.60-67, 1999.

[4] Changchien, S.W., and Lu, T.C., "Mining Association Rule Procedure to Support On-Line Recommendation by Customers and Products Fragmentation," *Expert Systems with Applications*, 20, pp.325-335, 2001.

[5] Hui, S.C. and Jha, G.., "Data Mining for Customer Service Support," *Information & Management*, 38, pp.1-13, 2000.

[6] Lee, K.C., Kim, J.S., Chung, N.H., and Kwon, S.J., "Fuzzy Cognitive Map Approach to Web-mining Inference Amplification," *Expert Systems with Applications*, 22, pp.197-211, 2002.

[7] Mitra, S. and Pal, S.K., "Logical Operation Based Fuzzy MLP for Classification and Rule Generation", *Neural Networks*, 7(2), pp.353-373. 1994.

[8] Song, H.S., Kim, J.K., and Kim, S.H., "Mining the Change of Customer Behavior in an Internet Shopping Mall," *Expert Systems with Applications*, 21, pp.157-168, 2001.